



UNIwersytet
Warszawski

Biuro Prasowe

26.10.2017

ARCHIWIZACJA WEBU

27 października z inspiracji UNESCO obchodzony jest Światowy Dzień Dziedzictwa Audiowizualnego. To okazja do zwrócenia uwagi na społeczne i naukowe znaczenie archiwów gromadzących zbiory audiowizualne, w tym także te, które posiadają wyłącznie postać cyfrową. Hasłem tegorocznych obchodów Światowego Dnia Dziedzictwa Audiowizualnego jest [„Odkrywać, pamiętać, upowszechniać”](#). Częścią dziedzictwa cyfrowego jest World Wide Web, najbardziej popularna i najłatwiej dostępna część Internetu. Archiwizacja Webu to jeden z tematów, którym zajmuje się Laboratorium Cyfrowe Humanistyki UW.

Publikowane powszechnie przynajmniej od połowy lat 90. strony WWW były i są archiwizowane w ramach projektów globalnych, takich jak archiwum prowadzone przez fundację Internet Archive, i krajowych – przez biblioteki czy archiwa narodowe. WWW sam w sobie nie ma żadnych mechanizmów archiwizacji stron i aktywne działanie w celu zabezpieczenia przynajmniej części jego zasobów jest konieczne. Wspomina o tym UNESCO w [dokumencie z 2015 roku](#), zalecając np. racjonalną archiwizację części zasobów mediów społecznościowych. Zalecenie to jest w praktyce realizowane m.in. przez brytyjskie Archiwa Narodowe w projekcie [UK Government Web Archive](#), gdzie zabezpieczone i udostępniane są np. archiwa tweetów z oficjalnych kont rządowych.

W Polsce archiwizacja zasobów WWW, w tym domeny krajowej .pl, nie jest prowadzona. Każda strona internetowa jest utworem w rozumieniu prawa autorskiego, a więc aby ją kopiować, taką kopię przechowywać i udostępniać należy mieć do tego odpowiednie uprawnienia. Tych wciąż nie mają polskie instytucje dziedzictwa.

W 2015 roku na Uniwersytecie Warszawskim powstało Laboratorium Cyfrowe Humanistyki. Ośrodek wspiera badaczy, którzy w swojej pracy już wykorzystują narzędzia cyfrowe lub chcą to robić. Udziela eksperckich porad przy planowaniu i realizowaniu projektów, organizuje warsztaty i seminaria, a także spotkania z przedstawicielami podobnych centrów z innych krajów. Jednym z tematów, którym zajmuje się LaCH UW, jest archiwizacja Webu.

– Nie wiemy, kiedy prawo autorskie umożliwi polskim instytucjom pamięci bezpieczne budowanie archiwum domeny krajowej – mówi Marcin Wilkowski z LaCH UW. – Tymczasem WWW jest dziś społecznie istotnym medium i dostęp do jego archiwów jest kluczowy dla wielu badań społecznych, językowych czy nawet dla postępowań prawnych. Czekając na zmiany w prawie, warto zdobywać i rozwijać kompetencje i metody związane z archiwizacją Webu i pracą z jego archiwami. To właśnie robimy w LaCH UW – tłumaczy.

Laboratorium Cyfrowe Humanistyki organizuje warsztat z zaawansowanych metod pracy z archiwum WWW udostępnianym przez fundację Internet Archive. Jest także współorganizatorem sesji i warsztatów poświęconych archiwizacji Internetu podczas konferencji „Toruńskie Konfrontacje Archiwalne”.

– Projekty archiwizacji WWW nigdy nie mają na celu zabezpieczenia wszystkich zasobów dostępnych online – mówi Marcin Wilkowski. – Wynika to nie tylko ze skali wielkości potencjalnych zbiorów czy przepisów prawa autorskiego, lecz także z ograniczeń samej metody. Archiwizacja Webu metodą web crawlingu pozwala na pobranie tylko wzajemnie linkujących się stron WWW i to na określoną „głębokość” takich powiązań. Do tego istnieją trudności w sczytywaniu stron wykorzystujących technologię Flash czy skrypty JavaScript. Co więcej, w przypadku archiwizacji prowadzonej przez Internet Archive każdy właściciel strony może blokować jej archiwizację przez umieszczenie odpowiednich wpisów w pliku robots.txt.